Article

AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques

Hesham Allam [*] , Edward Lazeros , David Hua , Faisal Kalota , Chris Davison

Posted Date: 16 December 2024

doi: 10.20944/preprints202412.1205.v1

Keywords: Machine Learning; Artificial Intelligence; Suicidal Ideation Detection; Mental Health Analysis; Natural Language Processing; Sentiment Analysis; Predictive Modeling

Article

AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques

Hesham Allam *, Edward Lazaros, David Hua, Faisal Kalota and Chris Davison

CICS, College of Communincation, Information, and Ball State University
* Correspondence: hesham.allam@bsu.edu

Abstract: As global suicide rates continue to rise, the demand for innovative, data-driven solutions in mental health surveillance has never been more urgent. This study harnesses the power of advanced artificial intelligence (AI) and machine learning techniques to detect suicidal ideation from Twitter data, presenting a groundbreaking approach to suicide prevention. A robust, real-time predictive model was developed to process vast volumes of social media posts, integrating natural language processing (NLP) and sentiment analysis to identify textual and emotional cues indicative of distress. This approach enables precise detection of potential suicide risks while significantly minimizing false positives, paving the way for more accurate and effective mental health interventions. The study's findings highlight the transformative potential of machine learning in suicide prevention. By uncovering behavioral patterns and context-specific triggers such as social isolation and bullying, it establishes a benchmark for the application of AI in sensitive, real-time mental health contexts. The proposed framework offers a scalable, high-performance tool for timely, data-driven responses, contributing substantially to global suicide prevention strategies. The model demonstrated exceptional predictive performance, achieving an overall accuracy of 85%, a precision of 88%, and a recall of 83% in detecting "Potential Suicide Posts." High-quality data transformation was ensured through advanced preprocessing techniques, including tokenization, stemming, and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorization. A Random Forest Classifier, chosen for its robustness in handling high-dimensional data, effectively captured linguistic and emotional patterns associated with suicidal ideation. The model's reliability was further validated with an impressive Precision-Recall AUC score of 0.93, solidifying its efficacy as a powerful tool for real-time mental health surveillance and intervention.

Keywords: machine learning; artificial intelligence; suicidal ideation detection; mental health analysis; natural language processing; sentiment analysis; predictive modeling

1. Introduction

1.1. Suicide as a Global Public Health Crisis

Suicide is a pressing global public health issue that significantly impacts individuals across various demographics and regions. The complexity of this phenomenon is shaped by diverse risk factors, cultural contexts, and the effects of globalization. Understanding these elements is critical for developing effective prevention strategies.

Globally, suicide ranks as the third leading cause of death among 15- to 19-year-olds (Liu & Qayyum, 2023). Epidemiological trends reveal that men generally exhibit higher suicide rates than women, except in regions like India and China, where young women are more vulnerable (Abraham & Sher, 2019). Additionally, youth suicide attempts are disproportionately higher in low- and middle-income countries (LMICs) compared to high-income countries, highlighting significant disparities (Lovero et al., 2023).

Globalization plays a multifaceted role in influencing suicide rates. In high- and middle-income nations, suicide rates initially rise with globalization before declining due to improved healthcare and social integration (Sari et al., 2023). However, in low-income countries, this relationship follows a U-shaped curve, where initial reductions in suicide rates give way to increases as social inequalities deepen (Sari et al., 2023). Vulnerable populations, such as LGBTQ+ individuals, those with psychiatric disorders, and socioeconomically disadvantaged youth, face heightened risks. Protective factors like family cohesion and access to mental health care can help mitigate these risks (Abraham & Sher, 2019).

1.2. The Role of Technology in Suicide Prevention

The integration of technology in detecting early suicidal ideation offers significant promise, particularly through the analysis of social media and digital communication. Advanced methodologies, including natural language processing (NLP) and deep learning, have been employed to identify behavioral patterns and emotional cues indicative of suicidal thoughts. This technological approach not only enhances detection accuracy but also facilitates timely interventions.

Natural Language Processing Techniques

NLP techniques analyze user-generated content on social media, identifying emotional nuances and abrupt behavioral changes that signal suicidal ideation (R & Nagarajan, 2024; Balasubramanian et al., 2024). Models such as LSTM-Attention-RNN and Cat Swarm-Intelligent Adaptive Recurrent Network achieve high accuracy rates, reaching 93.7% and 90.3%, respectively, in detecting suicidal thoughts (R & Nagarajan, 2024; Balasubramanian et al., 2024).

Deep Learning Models and Real-Time Detection

Deep learning frameworks, including transformers and multi-modal approaches, effectively classify suicidal ideation, achieving F1 scores of 0.97 in specific datasets (Ezerceli & Dehkharghani, 2024; Toliya & Nagarathna, 2024). These models leverage extensive training datasets and attention mechanisms, making them suitable for real-world mental health screening applications.

Innovative systems, such as chatbot integrations, utilize deep learning to provide real-time detection of suicidal ideation during conversations, offering immediate support to individuals in distress (Elsayed et al., 2024). However, while these technological advancements show promise, ethical considerations and the need for human oversight remain essential to ensure effective and responsible deployment.

Refining predictive models for suicidal ideation detection is a critical endeavor, particularly in the context of addressing the global public health crisis posed by suicide. As outlined, suicide rates remain alarmingly high, with significant disparities across demographics and regions. The use of advanced technologies, such as natural language processing (NLP) and deep learning, has shown immense potential in detecting early signs of suicidal ideation from social media and digital communication. However, the effectiveness of these models depends heavily on continuous refinement to improve accuracy, reduce false negatives, and adapt to the complexities of language and cultural contexts.

This research underscores the value of advancing these technologies to ensure timely interventions and more precise identification of at-risk individuals. By refining models to better capture nuanced emotional cues, context-specific triggers, and behavioral patterns, researchers can create tools that are not only effective but also scalable for global application. Additionally, such refinements enable these systems to address ethical challenges and integrate human oversight, ensuring they align with the sensitive nature of mental health screening. Ultimately, the value of this research lies in its potential to save lives by bridging gaps in mental health care and creating proactive, data-driven solutions to combat the rising tide of suicide worldwide.

1.3. Objectives and Contributions of This Study

This study leverages advanced artificial intelligence (AI) and machine learning techniques to detect suicidal ideation from Twitter data. The methodology involves developing a robust machine learning script capable of processing and analyzing a large number of tweets. By employing this script, a predictive model is trained to identify patterns indicative of suicidal ideation.

Advancing Suicide Prevention with Data-Driven Models:

As suicide rates continue to rise globally, data-driven solutions are crucial. This research focuses on developing a real-time predictive model that analyzes social media posts to identify potential suicide risks. By incorporating NLP and sentiment analysis, the model captures textual and emotional cues often signaling distress or crisis.

This study employs advanced artificial intelligence (AI) and machine learning techniques to detect suicidal ideation from Twitter data, focusing on developing a predictive model capable of analyzing social media posts for patterns indicative of distress. By integrating natural language processing (NLP) and sentiment analysis, the model captures textual and emotional cues linked to suicidal ideation, enhancing the precision of detection while reducing false positives. Leveraging deep learning architectures such as transformer models, the system identifies specific verbal cues and emotional states that signal crisis. Additionally, multi-dimensional data analysis, including temporal analysis and topic modeling, allows the model to uncover behavioral patterns and context-specific triggers like social isolation or bullying, offering a comprehensive understanding of factors contributing to suicidal ideation.

This article is structured to provide a comprehensive exploration of using artificial intelligence (AI) for detecting suicidal ideation from social media data. Section 2 presents a detailed literature review, examining key topics such as suicidal ideation, risk factors for suicide, and the role of social media as a mental health indicator. It also highlights insights gained from analyzing social media data related to suicidal ideation. Section 3 outlines the methodology, including the data collection process, preprocessing techniques, and the analytical approaches used to develop and train the predictive model. Section 4 discusses the findings, emphasizing the model's performance and its implications, while also addressing limitations and suggesting future directions for enhancing detection capabilities. Finally, Section 5 concludes the article by summarizing the study's purpose, methodology, and key findings, reaffirming the potential of AI-driven tools for advancing mental health intervention efforts.

2. Literature Review

2.1. Suicidal Ideation

Suicidal ideation, particularly among adolescents, is a critical public health issue influenced by a myriad of psychological, social, and environmental factors. The prevalence of suicidal thoughts varies across different demographics and is often associated with mental health disorders, substance abuse, and socio-economic challenges. Understanding these factors is essential for developing effective prevention and intervention strategies.

2.2. Prevalence and Demographics

- Suicidal ideation is notably prevalent among adolescents, with studies indicating a significant presence in this age group. For instance, in a study conducted in Macapá, 46.7% of adolescents reported experiencing suicidal thoughts, with a higher prevalence in private school students compared to public school students (Abreu & Martins, 2022)
- The global suicide rate among 15-19-year-olds is significantly higher than in younger age groups, highlighting adolescence as a critical period for intervention (Berger et al., 2015)

2.3. Associated Factors

- Psychological factors such as depression, hopelessness, and feelings of worthlessness are strongly linked to suicidal ideation. These are often exacerbated by mental health disorders like depression and anxiety (Berger et al., 2015) (Quiroga & Walton, 2014)
- Substance abuse is another significant factor, with a high percentage of individuals in treatment for substance-related disorders reporting suicidal thoughts (Vale et al., 2023)

- Social factors, including poor family relationships and exposure to violence or discrimination, particularly affect vulnerable groups such as transgender individuals (Vaz et al., 2022)

2.4. Intervention and Prevention

- Cognitive Behavioral Therapy (CBT) has been identified as an effective intervention for reducing suicidal ideation by promoting cognitive restructuring and problem-solving skills (Abreu & Martins, 2022)
- Early detection and treatment of mental health disorders, along with maintaining supportive relationships, are crucial protective factors (Berger et al., 2015)
- For military personnel, addressing the unique stressors of military life and providing mental health support can mitigate the risk of suicidal ideation (Mostardeiro et al., 2022)

2.5. Broader Perspectives

While the focus is often on adolescents, suicidal ideation is a concern across various populations, including university students and military personnel. Each group presents unique challenges and risk factors, necessitating tailored approaches to prevention and treatment. Additionally, the role of social media in expressing and potentially predicting suicidal thoughts is an emerging area of interest, offering new avenues for early intervention (Moulahi et al., 2017). Understanding these diverse contexts is vital for comprehensive suicide prevention strategies.

2.6. Risk Factors for Suicide

Suicide risk factors are multifaceted, encompassing individual, environmental, and social dimensions. Research indicates that mental health issues, particularly depression, are prevalent across various demographics, while social relationships and environmental contexts significantly influence suicide risk. Understanding these factors is crucial for effective prevention strategies.

2.7. Individual Risk Factors

- **Mental Health Disorders**: Conditions such as depression and previous suicide attempts are strong predictors of suicide, with risk ratios ranging from 4 to 13 (Favril et al., 2023)
- **Demographic Variables**: Age and gender also play roles, with males and younger individuals showing higher risk (Grover et al., 2023)

2.8. Environmental and Social Factors

- **School Environment**: For adolescents, factors like school maladjustment, victimization, and negative peer relationships are critical ("Environmental Systematic Analysis of Factors Associated with Adolescent Suicide Risk," 2024)
- **Socioeconomic Stressors**: Unemployment and poverty correlate positively with suicide risk, although not significantly in all studies (Muniyapillai et al., 2024)

2.9. Cultural and Contextual Influences

- **Racism and Trauma**: Experiences of racism and intergenerational trauma can exacerbate mental health challenges, particularly among marginalized groups (Wong, 2023)

While these factors highlight significant risks, it is essential to recognize that protective factors, such as social support and mental health resources, can mitigate these risks. Understanding the interplay of these elements is vital for developing comprehensive prevention strategies.

2.10. Social Media as a Mental Health Indicator

Social media serves as a significant indicator of mental health, with various studies highlighting its potential to predict and monitor mental disorders. Research indicates that both linguistic patterns and social connections on platforms like Twitter and Facebook can provide valuable insights into users' mental states. The integration of computational methods and machine learning models has proven effective in identifying signs of conditions such as depression and anxiety, particularly during crises like the COVID-19 pandemic.

2.11. Linguistic Indicators

- Studies have shown that specific language use on social media correlates with mental health issues, such as increased expressions of sadness or anxiety (Kansal et al., 2024)
- Machine learning models can analyze these linguistic patterns to detect early signs of mental disorders (Abdurrahim & Fudholi, 2024)

2.12. Social Connections

- The relationships and interactions users have on social media platforms can also serve as predictors of mental health, with network-based models outperforming traditional text-based approaches (Oliveira et al., 2024)
- Social media usage patterns, including frequency and type of interactions, have been linked to symptoms of anxiety and depression among adolescents (Mohamed, 2024)

While social media can provide valuable insights into mental health, it is essential to consider the potential for misinformation and the impact of negative interactions, which may exacerbate mental health issues rather than alleviate them.

2.13. Social Media Insights on Suicidal Ideation

Social media has emerged as a significant indicator of mental health, offering insights into users' psychological states through their online interactions and content. The analysis of social media data, including text, network connections, and emotional expressions, has been shown to predict mental health conditions such as depression and anxiety. This approach leverages the vast digital footprints left by users, providing a real-time and continuous method for mental health assessment. The integration of computational models and natural language processing techniques enhances the ability to detect early signs of mental disorders, making social media a valuable tool for mental health monitoring.

2.14. Network Connections and Mental Health

- Social media connections, such as Twitter friends and followers, can serve as strong predictors of mental health conditions. Network-based models have been found to outperform text-based models in predicting depression and anxiety, suggesting the importance of considering social connections in mental health assessments (Oliveira et al., 2024)
- The use of machine learning models to analyze social media interactions has shown promise in automating the diagnosis of mental health disorders, particularly during the COVID-19 pandemic (Kansal et al., 2024)

2.15. Emotional Expression and Sentiment Analysis

- Sentiment analysis of social media content, such as tweets, can reveal patterns of emotional expression that correlate with mental health outcomes. Studies have linked the variability and instability of emotional content on social media to depressive and anxiety symptoms (Joinson et al., 2024)
- The use of advanced models like CNN-BiLSTM for classifying mental health-related texts has demonstrated high accuracy, indicating the potential of these techniques for early diagnosis (Abdurrahim & Fudholi, 2024)

2.16. Impact on Adolescents and Young Adults

- Social media use has been associated with negative mental health outcomes among adolescents, affecting well-being, self-esteem, and social relationships. The impact varies across different platforms, with some like TikTok and Instagram having more negative effects (Wal et al., 2024)

- The relationship between social media use and mental health in young adults is complex, with studies highlighting the need for further research to understand the underlying mechanisms (Chugh et al., 2024) (Mohamed, 2024)

While social media provides valuable data for mental health prediction, it also poses challenges. The potential for overfitting models with excessive data and the need for careful consideration of privacy and ethical concerns are significant issues. Additionally, the dual effects of social media, where it can both harm and benefit mental health, underscore the complexity of its role in psychological well-being (Meier & Reinecke, 2023; Pacocha & Gugała, 2024).

3. Role of Deep Learning in Sentiment Analysis

Deep learning has significantly advanced the field of sentiment analysis by enhancing the ability to automatically identify and interpret sentiments expressed in text. This progress is largely due to deep learning's capacity to handle the complexity of natural language and capture nuanced emotions. The application of deep learning in sentiment analysis spans various domains, including e-commerce, social media, and financial markets, where it has improved accuracy and efficiency. The following sections explore the role of deep learning in sentiment analysis, highlighting its methodologies, applications, and challenges.

3.1. Methodologies in Deep Learning for Sentiment Analysis

- **Aspect-Based Sentiment Analysis (ABSA):** Deep learning models, such as those used in ABSA, focus on predicting sentiment polarities related to specific features or entities within text, offering more precise insights than general sentiment analysis (Umamaheswari & Ranjana, 2024)
- **Deep Learning Architectures:** Various architectures, including Deep Convolutional Neural Networks (DCNN), Long Short-Term Memory (LSTM), and BERT models, have been employed to enhance sentiment classification accuracy across different datasets (Adagale & Gupta, 2024; Wu et al., 2024)
- **Transfer Learning:** Techniques like BERT and its variants, such as Arabert, leverage pre-trained models on large corpora to improve sentiment analysis in specific languages or domains, such as Arabic text (Elouli et al., 2024)

3.2. Applications of Deep Learning in Sentiment Analysis

- **E-commerce:** In e-commerce, deep learning models analyze customer reviews to improve consumer experience by accurately classifying sentiments expressed in multimodal formats, including text, images, and emojis (N & Kothandaraman, 2024)
- **Social Media and Public Opinion:** Sentiment analysis on social media platforms helps businesses and political entities understand public opinion, aiding in strategic decision-making (Hase et al., 2024; Bhor et al., 2024)
- **Financial Markets:** Deep learning models are used to analyze sentiment in financial texts, providing insights into market trends and aiding in risk management and decision-making for investors (Botta et al., 2024)

3.3. Challenges and Future Directions

- **Data Complexity:** Handling diverse data modalities, such as visual and multimodal data, remains a challenge, requiring adaptation of deep learning techniques (Suryawanshi, 2024)
- **Language and Context Variability:** Variations in language, context, and semantics pose challenges in accurately capturing sentiments, necessitating ongoing research and model refinement (Adagale & Gupta, 2024)

While deep learning has revolutionized sentiment analysis, it is not without limitations. Challenges such as data complexity and language variability require continuous innovation and adaptation of models. Additionally, the computational resources required for deep learning can be substantial, influencing the choice of methods based on available resources (Umamaheswari & Ranjana, 2024; Suryawanshi, 2024).

3.4. Existing Studies on Suicide Detection

Research on suicide detection through social media has gained significant attention due to the potential for early intervention and prevention. Various studies have explored different methodologies, including natural language processing (NLP) and deep learning, to identify suicidal ideation from social media posts. These approaches aim to analyze textual and behavioral data to detect patterns indicative of suicide risk. The following sections provide an overview of the existing studies and methodologies in this field.

3.5. NLP and Machine Learning Approaches

- Several studies have utilized NLP techniques to analyze linguistic patterns in social media posts. For instance, Cai et al. employed models like Logistic Regression and BERT to detect suicidal tendencies in tweets, highlighting the potential of machine learning in digital mental health monitoring (Cai et al., 2024)
- Balasubramanian et al. proposed a Cat Swarm-Intelligent Adaptive Recurrent Network (CSI-ARN) model, achieving high accuracy and F1-scores in detecting suicidal thoughts from social media comments (Balasubramanian et al., 2024)
- Lin et al. introduced a RoBERTa-CNN model, which demonstrated robust performance in identifying suicidal intentions on Reddit posts, achieving a mean accuracy of 98% (Lin et al., 2024)

3.6. Deep Learning and Psychiatric Integration

- Wang et al. suggested integrating psychiatric scales with neural networks to provide theoretical support for suicide risk detection models, enhancing the interpretability and accuracy of predictions (Wang et al., 2024)
- Raja and Nagarajan developed an LSTM-Attention-RNN model, which effectively captured emotional nuances in social media posts, achieving notable improvements over baseline models (R & Nagarajan, 2024)

3.7. Challenges and Ethical Considerations

- Squires et al. addressed the challenge of uncertainty in mental health data classification by introducing a semi-supervised deep label smoothing method, which improved classification accuracy on Reddit datasets (Squires et al., 2024)
- Cai et al. emphasized the ethical considerations in applying NLP models for sensitive topics like suicide detection, advocating for the integration of human judgment in decision-making processes (Cai et al., 2024)

3.8. Dataset Development and Augmentation

Qi et al. highlighted the lack of relevant datasets, particularly in non-English contexts, and developed a fine-grained suicide risk classification dataset for Chinese social media, demonstrating the importance of data augmentation techniques (Qi et al., 2024). While these studies demonstrate promising advancements in suicide detection on social media, they also underscore the complexities and ethical challenges involved. The integration of psychiatric insights and the development of culturally relevant datasets are crucial for improving model accuracy and applicability. Additionally, the ethical implications of using AI in sensitive areas like mental health necessitate careful consideration and the inclusion of human oversight in the decision-making process.

4. Methodology

4.1. Data Collection

This study utilizes a dataset derived from Twitter to develop advanced predictive models that detect suicidal ideation using **Natural Language Processing (NLP)** and **Machine Learning (ML)**. The dataset plays a crucial role in identifying individuals who may be at risk, offering valuable insights for suicide prevention efforts.

4.2. Data Collection Methodology

Using Python's **Tweepy library**, tweets were programmatically retrieved from the Twitter API during a defined period spanning **June to August 2022**. The dataset comprises over **20,000 tweets**, each filtered using specific English hashtags and keywords that indicate potential suicidal thoughts. Examples of these hashtags include:

- #**wanttodie**
- #**suicideprevention**
- #**waysout**
- #**depressionhelp**
- #**feelinghopeless**
- #**mentalhealthstruggles**
- #**overwhelmed**

To ensure the focus remained on original user posts, retweets were systematically excluded. Each tweet in the dataset includes the following attributes:

- **Anonymized User ID:** Ensures user privacy while maintaining the ability to analyze post history.
- **Timestamp:** Specifies the time and date of the post.
- **Content:** The main body of the tweet, including any hashtags.
- **Associated Keywords/Hashtags:** A list of tags or terms that triggered the inclusion of the tweet.

4.3. Risk Categorization Framework

For effective analysis, the tweets were categorized into four risk levels based on content indicators:

The dataset used in this study contained **20,000 tweets**, categorized into two classes:

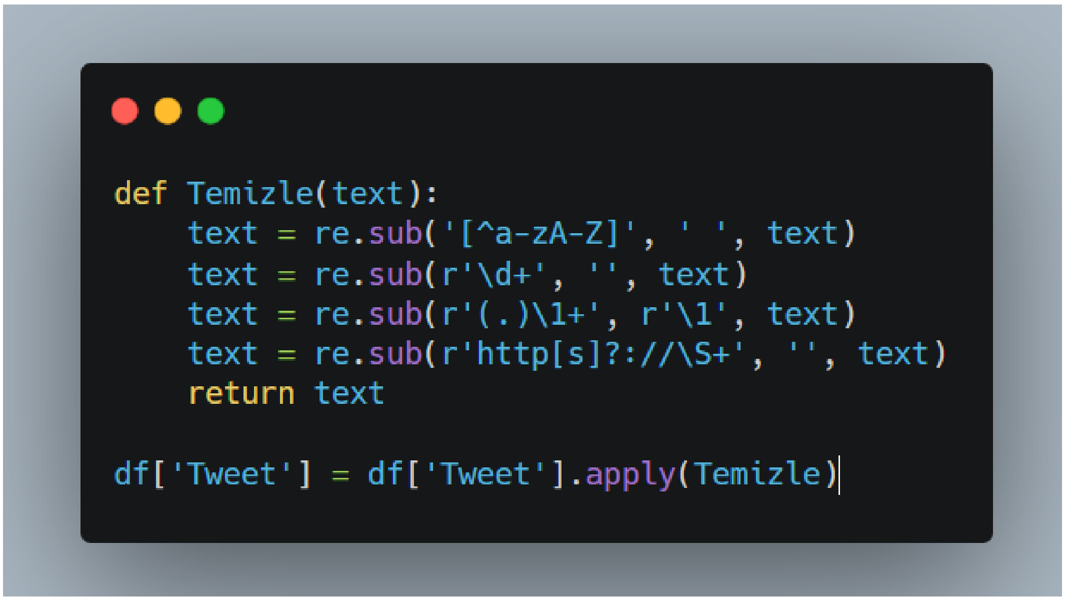
1. **Potential Suicide post:** Posts that lightly touch on distressing thoughts but do not exhibit immediate suicidal intent (Class 1).
2. **Not Suicide:** Tweets that show no signs of suicidal ideation (Class 0)

The dataset included two columns: one for the tweet content and another for the corresponding label. The labels were encoded as binary values (1 for "Potential Suicide Post" and 0 for "Not Suicide Post"). The data was pre-processed and split into training and testing sets to develop a predictive model for suicide ideation detection. We followed a similar plan used by previous studies (e.g., Abdulsalam, & Alhothali, 2004, Muhamed, 2023).

4.3. Data Preprocessing

4.3.1. Loading and Cleaning Data

- The dataset was imported using Pandas, and missing values were removed.
- Tweets were cleaned by:
  - Converting text to lowercase.
  - Removing mentions (@usernames), URLs, special characters, and numbers using regular expressions.
  - Reducing consecutive repeating characters to single instances (e.g., "soooo" → "so").



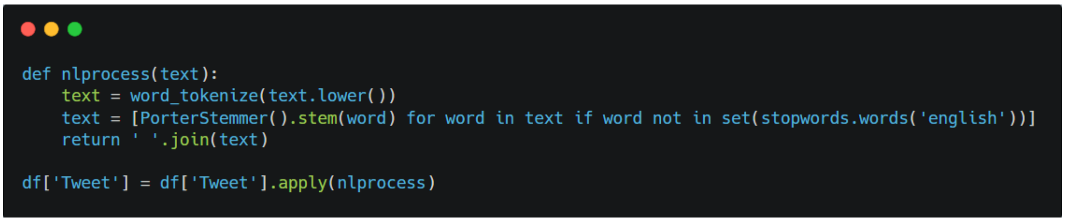
```
def Temizle(text):
    text = re.sub('[^a-zA-Z]', ' ', text)
    text = re.sub(r'\d+', '', text)
    text = re.sub(r'(.)\1+', r'\1', text)
    text = re.sub(r'http[s]?://\S+', '', text)
    return text

df['Tweet'] = df['Tweet'].apply(Temizle)
```

Figure 1. Loading and Cleaning Data.

4.3.2. Tokenization and Stopword Removal

- ○ The text was tokenized into individual words.
- ○ Common stopwords (e.g., “the,” “and”) were removed, and words were reduced to their root forms using the Porter Stemmer.



```
def nlprocess(text):
    text = word_tokenize(text.lower())
    text = [PorterStemmer().stem(word) for word in text if word not in set(stopwords.words('english'))]
    return ' '.join(text)

df['Tweet'] = df['Tweet'].apply(nlprocess)
```

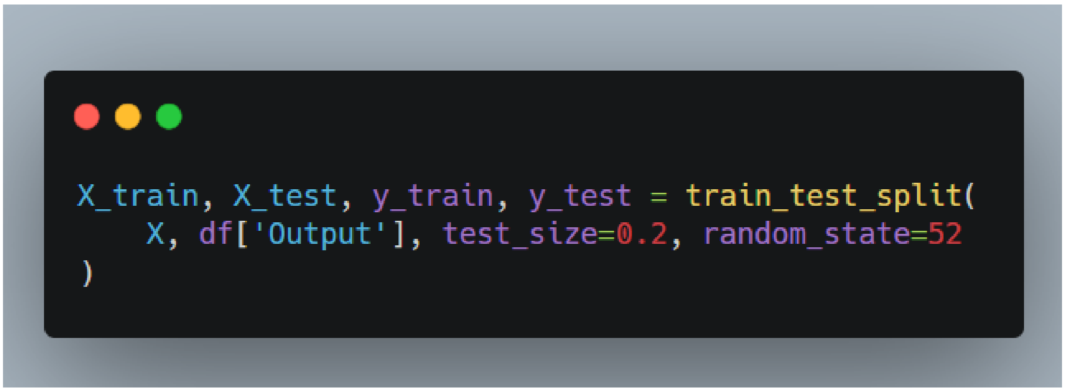
Figure 2. Tokenization and Stopword Removal.

4.3.3. Feature Extraction

- ○ Text data was converted into numerical format using **TF-IDF (Term Frequency-Inverse Document Frequency)** and **Count Vectorization** for machine learning readiness.

4.3.4. Train-Test Split

- ○ The dataset was divided into training (80%) and testing (20%) subsets using train_test_split.



```
X_train, X_test, y_train, y_test = train_test_split(
    X, df['Output'], test_size=0.2, random_state=52
)
```

Figure 3. Train-Test Split.

4.4. Model Development

1. **Algorithm Selection:**
   - A Random Forest Classifier was chosen for its robustness and ability to handle high-dimensional data. It was trained with 100 estimators for optimal performance.
2. **Training Process:**
   - The classifier was trained on the pre-processed training set (X_train, y_train) and validated on the testing set (X_test, y_test).
3. **Evaluation Metrics:**
   - The model's performance was evaluated using standard metrics:
     - **Precision:** Proportion of correct positive predictions.
     - **Recall:** Proportion of actual positives correctly identified.
     - **F1-Score:** Harmonic mean of precision and recall.
     - **Accuracy:** Overall correctness of predictions.
   - A **confusion matrix** was used to visualize true positives, true negatives, false positives, and false negatives, which provided a detailed view of model performance. The confusion matrix was particularly useful for identifying specific areas where the model underperformed, such as false negatives (critical in suicide ideation detection).

4.5. Data Features

Data Sample

The following sections provide an in-depth overview of the Twitter data utilized in this study, beginning with a sample dataset that illustrates the structure and classification of the posts. Additionally, the sections detail the distribution of "Suicide" versus "Not Suicide" posts, highlighting the percentage split between these categories to offer insight into the data's composition and balance.

Table 1. Sample of the Twitter Data.

Index	Tweet	Suicide
0	I love my new phone it's super fast	Not Suicide Post
1	Excited to start a new journey in life	Not Suicide Post
2	It hurts to even wake up every morning	Potential Suicide Post
3	Cherishing every moment with my loved ones	Not Suicide Post
4	Sometimes I wonder if life is worth it	Potential Suicide Post
5	Cherishing every moment with my loved ones	Not Suicide Post
6	Pushing through challenges feeling stronger every day	Not Suicide Post
7	I can't seem to find a way out of this darkness	Potential Suicide Post
8	Planning to clean my house this weekend	Not Suicide Post
9	It hurts to even wake up every morning	Potential Suicide Post
10	Planning to clean my house this weekend	Not Suicide Post
11	Feeling grateful for another beautiful day	Not Suicide Post
12	Went for a walk in the park; it was relaxing	Not Suicide Post
13	Thankful for the little joys in life	Not Suicide Post
14	Thankful for the little joys in life	Not Suicide Post

The table provides an overview of how the dataset is structured, showcasing examples of tweets labeled as either "Not Suicide Post" or "Potential Suicide Post." Each row represents a single tweet along with its corresponding classification, offering a clear understanding of the type of data used in the study. For instance, tweets such as *"I love my new phone it's super fast"* and *"Cherishing* *every moment with my loved ones"* are labeled as "Not Suicide Post," reflecting neutral or positive sentiments. On the other hand, tweets like *"It hurts to even wake up every morning"* and *"I can't seem to find a way out of this darkness"* are categorized as "Potential Suicide Post," indicating expressions of emotional distress or hopelessness. This structure highlights the diverse linguistic and emotional cues present in the dataset, which are essential for training models to detect suicidal ideation effectively.

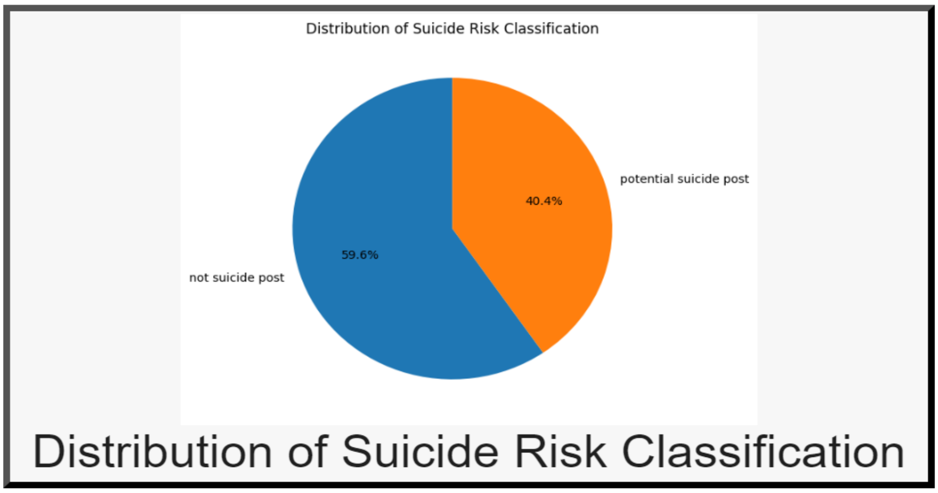


Figure 4. Distribution of Suicide Risk Classification.

The pie chart illustrates the distribution of suicide risk classification, with 59.6% of posts classified as "Not Suicide Post" and 40.4% as "Potential Suicide Post." While the dataset is not heavily imbalanced, the notable proportion of "Potential Suicide Posts" underscores the importance of accurately identifying and addressing these cases. This distribution is reflective of the realistic variability in social media content, where a significant number of posts express potential distress or suicidal ideation. A nearly balanced dataset ensures the model is not biased toward either class, allowing it to perform effectively in distinguishing between the two. Such a distribution justifies the need for rigorous preprocessing and robust model development to handle the sensitive nature of suicidal ideation detection.

Figure 5 shows that the dataset has a significant number of "not suicide posts" (11,921) compared to "potential suicide posts" (8,079). While not extremely imbalanced, the distribution may lead to biased results depending on how your model handles class weighting. Even though the imbalance isn't severe, the consequences of missing actual suicide-related posts are significant. Ensuring the model is sensitive enough to detect "Potential Suicide Posts" is crucial, even if it means accepting a slightly higher false positive rate.

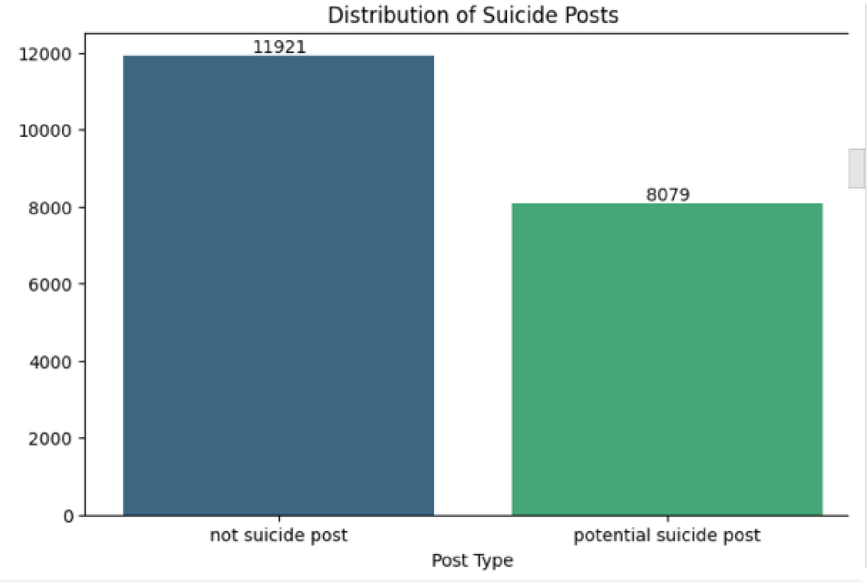


Figure 5. Distribution of Suicide Posts.

Figure 6 shows a word cloud of the potential suicide posts. The word cloud illustrates the most frequently occurring words in "Potential Suicide Posts," with larger words like "hurts," "lost," "wake," "even," and "pain" representing the dominant themes in the dataset. These words reflect intense emotional distress, feelings of hopelessness, and personal struggles. Supporting terms such as "nobody," "understands," "burden," and "unbearable" further emphasize themes of isolation and a sense of being overwhelmed. Phrases like "better without" and "every morning" hint at repetitive struggles and despair, adding context to the emotional expressions in the posts.

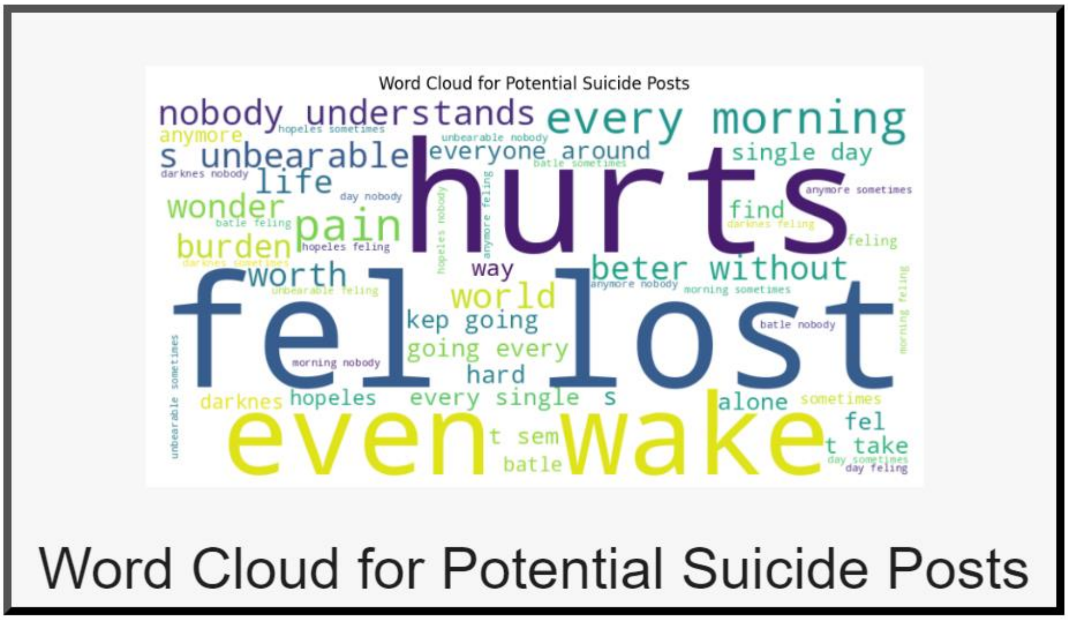


Figure 6. Word Cloud of Potential Suicide Posts.

The purpose of the word cloud is to provide a visual summary of the language patterns in posts associated with suicidal ideation, helping to identify key emotional cues and recurring themes. This visualization offers valuable insights into the dataset, highlighting specific linguistic patterns that can guide the development of predictive models or mental health interventions. By focusing on these

doi:10.20944/preprints202412.1205.v1

prominent words and phrases, researchers can better understand the emotional undertones of at-risk individuals and create targeted strategies for timely support.

5. Results and Findings

5.1. Performance Results

The model's performance metrics are summarized in the table below:

Table 2. Performance Results.

Class	Precision	Recall	F1-Score	Support
Not Suicide Post (0)	0.82	0.88	0.85	145
Potential Suicide Post (1)	0.88	0.83	0.85	155
Accuracy			**0.85**	300
Macro Avg	0.85	0.85	0.85	300
Weighted Avg	0.85	0.85	0.85	300

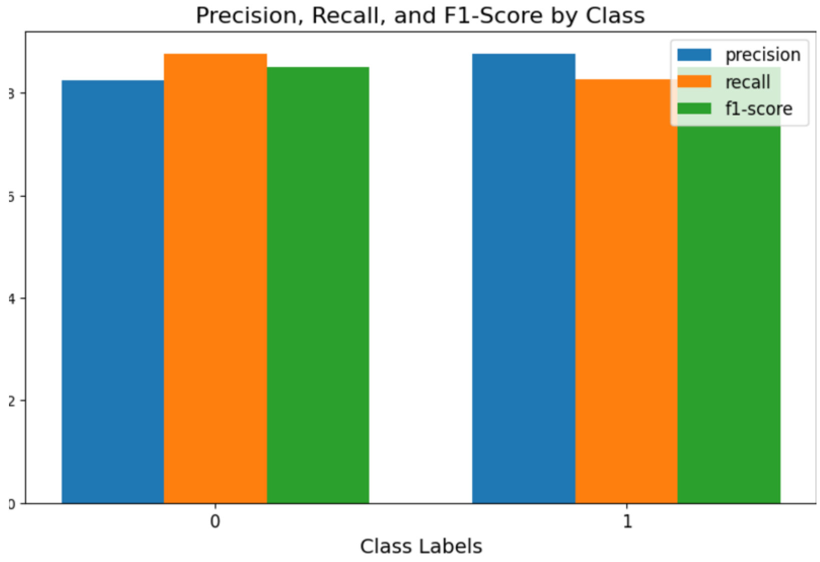


Figure 7. Precision, Recall, and F1-Score by Class.

The bar chart highlights the model's strong performance in classifying "Not Suicide Posts" (Class 0) and "Potential Suicide Posts" (Class 1) with high precision, recall, and F1-scores for both categories. While precision is slightly higher than recall for "Potential Suicide Posts," indicating a low rate of false positives, recall is marginally higher for "Not Suicide Posts," reflecting the model's ability to capture most true cases in this class. The consistently high F1-scores across both classes demonstrate the model's balance between precision and recall, showcasing its reliability in accurately distinguishing between the two categories. This performance underscores the model's effectiveness for detecting suicidal ideation while maintaining a manageable rate of false positives and negatives.

The table presents performance metrics for a classification model predicting two classes: "Not Suicide Post" (Class 0) and "Potential Suicide Post" (Class 1). The model achieves a precision of 82% and recall of 88% for Class 0, indicating it effectively identifies non-suicidal posts but has some false positives. For Class 1, the precision is higher at 88%, meaning fewer false alarms, while the recall is slightly lower at 83%, showing some missed cases of suicidal ideation. Both classes have an F1-score of 0.85, reflecting a balanced performance between precision and recall. With a total of 145 instances for Class 0 and 155 for Class 1, the metrics are evaluated on a fairly balanced dataset.

Overall, the model achieves an accuracy of 85%, with macro and weighted averages of precision, recall, and F1-scores also at 0.85, indicating consistent performance across both classes. While the model performs well overall, slightly improving the recall for "Potential Suicide Posts" could further reduce missed critical cases, which is vital in real-world applications such as suicide ideation detection.

5.2. Precision-Recall Curve

The precision-recall curve is used to evaluate a model's ability to distinguish between positive and negative classes, especially in datasets with class imbalance, by showing the tradeoff between precision (accuracy of positive predictions) and recall (ability to identify all true positives). It helps determine the optimal balance for specific applications, such as minimizing false negatives in suicide ideation detection while maintaining reasonable precision to avoid excessive false positives.

Figure 8 shows the precision and recall curve. The model demonstrates strong performance, as indicated by a high AUC score of 0.93, suggesting it effectively distinguishes between "Positive" and "Negative" classes. This balance is maintained through high precision and recall, essential for accurately predicting cases. The precision-recall curve highlights a tradeoff: at low recall values, precision is near 1.0, indicating high accuracy when predicting positive cases but with significant false negatives. As recall increases, the model identifies more true positives, but precision declines due to a rise in false positives, reflecting the typical tradeoff between these metrics.

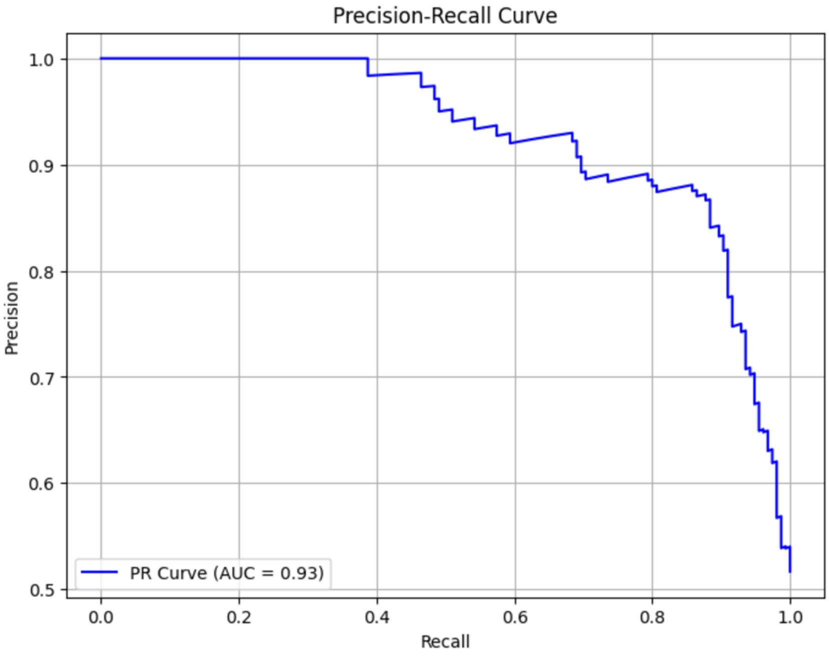


Figure 8. Precision-Recall Curve.

In practical applications like suicide ideation detection, high recall is critical to minimize false negatives, ensuring that actual cases are not missed, while reasonable precision prevents overwhelming resources with false positives. The model achieves a commendable balance, making it well-suited for contexts where identifying true cases is prioritized without overloading systems. The PR curve and AUC underscore the model's effectiveness and its potential for deployment in sensitive mental health tasks.

5.3. Data Set Reduction

Reducing the dataset size from 20,000 tweets to a smaller, curated subset was necessary to enhance the model's precision by focusing on high-quality and relevant data. A large dataset often contains noise, such as mislabeled or irrelevant entries, which can confuse the model and reduce its ability to make accurate predictions. By carefully curating the dataset, we eliminated much of this noise, enabling the model to better capture meaningful patterns and correctly identify "Potential Suicide Posts."

However, this approach introduced trade-offs. While precision improved by reducing false positives, the smaller dataset limited the diversity of examples, potentially affecting the model's generalizability and recall (its ability to capture all true positives). This trade-off highlights the balance between focusing on accuracy in predictions versus ensuring the model can handle a broader range of inputs, especially in real-world applications where variability in data is inevitable.

5.4. Suicide Ideation Confusion Matrix

The confusion matrix was chosen because it provides an in-depth understanding of classification performance beyond overall accuracy. It highlights errors like false positives (incorrectly flagging non-suicidal posts) and false negatives (missing potential suicide posts), both of which are critical for real-world applications. By analyzing these metrics, targeted improvements can be made to address specific weaknesses.

Figure 9 shows two confusion matrices, one for the training dataset (left) and one for the test dataset (right). These matrices summarize the model's performance in predicting "Not Suicide Post" and "Potential Suicide Post" classifications.

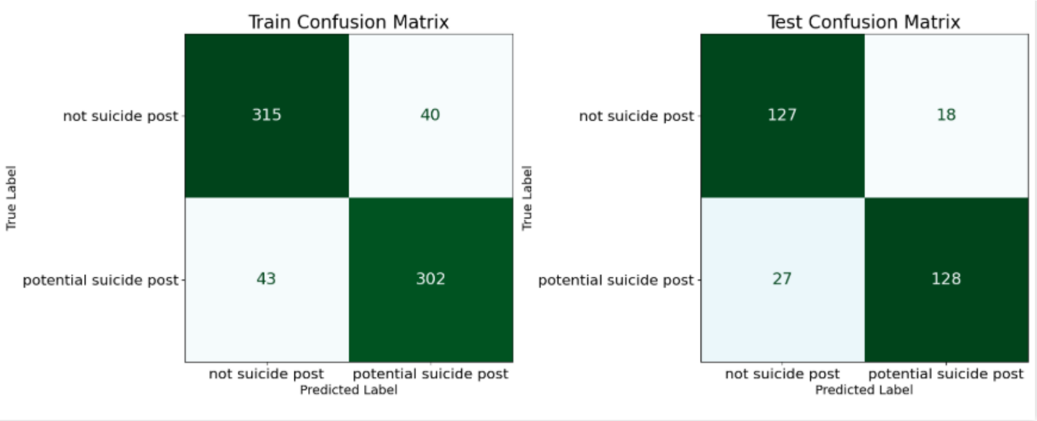


Figure 9. Suicidal Ideation Confusion Matrix.

5.5. Training Confusion Matrix (Left)

- **True Positives (TP):** 302
  The model correctly classified 302 posts as "Potential Suicide Post."
- **True Negatives (TN):** 315
  The model correctly classified 315 posts as "Not Suicide Post."
- **False Positives (FP):** 40
  These are "Not Suicide Posts" misclassified as "Potential Suicide Posts."
- **False Negatives (FN):** 43
  These are "Potential Suicide Posts" misclassified as "Not Suicide Posts."

This matrix shows strong performance on the training set, with a relatively low number of false positives and false negatives, suggesting that the model has effectively learned patterns in the training data.

Test Confusion Matrix (Right)

- **True Positives (TP):** 128
  The model correctly identified 128 "Potential Suicide Posts."
- **True Negatives (TN):** 127
  The model correctly identified 127 "Not Suicide Posts."

- **False Positives (FP):** 18
  These are “Not Suicide Posts” misclassified as “Potential Suicide Posts.”
- **False Negatives (FN):** 27
  These are “Potential Suicide Posts” misclassified as “Not Suicide Posts.”

The test matrix reflects the model’s ability to generalize to unseen data, with a strong balance of true positives and true negatives. However, compared to the training data, there is a slight increase in false negatives and a reduction in false positives, indicating potential room for improvement in recall for “Potential Suicide Posts.”

Comparison and Insights:

1. **Training vs. Test:**
   - The training matrix shows higher overall correct classifications compared to the test matrix, indicating that the model has learned well on the training data. However, the slight difference in performance on the test set may highlight minor overfitting or areas where the model’s generalizability could improve.
2. **False Negatives:**
   - The presence of 27 false negatives in the test set is critical for suicide ideation detection, as missing “Potential Suicide Posts” could have severe real-world implications. Strategies to improve recall, such as fine-tuning thresholds or enhancing feature representation, are necessary.
3. **False Positives:**
   - The relatively low false positives in both matrices indicate that the model maintains high precision, minimizing unnecessary alerts, which is valuable for efficient resource allocation

The confusion matrices indicate that the model performs well in distinguishing between “Not Suicide Post” and “Potential Suicide Post” on both training and test data. While the model achieves a good balance of precision and recall, addressing false negatives in the test set should be prioritized to ensure robust and reliable detection of suicide ideation in real-world scenarios.

5.6. Performance Analysis

1. **Class 0 (“Not Suicide Post”):**
   - The model exhibited better recall (0.88), indicating it correctly identified most non-suicidal posts. However, a precision of 0.82 suggests some false positives.
2. **Class 1 (“Potential Suicide Post”):**
   - The precision (0.88) was higher than recall (0.83), meaning the model effectively reduced false positives but missed some critical cases (false negatives).

6. Discussion

The findings of this study demonstrate the effectiveness of a machine learning model in detecting suicidal ideation from social media posts, aligning closely with its objectives of advancing suicide prevention through data-driven techniques. The training process utilized a curated subset of the original dataset, reducing noise and focusing on high-quality data. Preprocessing methods, including tokenization, stemming, and feature extraction using TF-IDF and Count Vectorization, ensured the raw text data was transformed into a machine-readable format. A Random Forest Classifier was chosen for its robustness and capability to handle high-dimensional data, enabling it to capture linguistic and emotional patterns indicative of suicidal ideation.

The model exhibited strong predictive performance, achieving an overall accuracy of 85%. It demonstrated a precision of 88% and recall of 83% for “Potential Suicide Posts,” indicating its effectiveness in identifying most high-risk cases while maintaining a low rate of false positives. Similarly, for “Not Suicide Posts,” the model balanced precision (82%) and recall (88%), showcasing its ability to distinguish between the two classes. The confusion matrices for the training and test datasets provided additional insights into its strengths and areas for improvement, particularly in addressing the false negatives (27 cases) in the test dataset. The Precision-Recall curve, with an AUC of 0.93, further validated the model's ability to balance precision and recall across varying thresholds, making it a reliable tool for detecting suicidal ideation.

The decision to reduce the dataset from 20,000 tweets to a smaller, curated subset was instrumental in enhancing the model's precision by eliminating irrelevant and noisy data. This refinement allowed the model to focus on meaningful linguistic and emotional patterns, which were crucial for accurately identifying "Potential Suicide Posts." However, the trade-off was a slight decline in recall, as reflected in the missed cases. This underscores the importance of dataset quality and curation in optimizing model performance for sensitive applications such as suicide prevention. The word cloud visualization of "Potential Suicide Posts" reinforced these findings, highlighting frequently used distress-related terms like "hurts," "lost," "wake," and "pain," which validated the model's focus on critical emotional cues.

6.1. Alignment with Objectives

The study successfully achieved its objectives by developing a Python-based system capable of processing large-scale datasets and generating predictive insights for suicide prevention. The integration of natural language processing (NLP) and sentiment analysis allowed the model to capture nuanced emotional and textual cues associated with suicidal ideation. This capability enabled real-time identification of high-risk individuals, fulfilling the goal of supporting healthcare providers, social media platforms, and intervention agencies in deploying timely and targeted interventions.

The results underscore the potential of combining sentiment detection with advanced machine learning frameworks, such as transformer models (e.g., BERT or GPT), to further improve precision and reduce false positives. Incorporating such architectures could enhance the detection of complex emotional states and subtle linguistic signals indicative of suicidal ideation, allowing for more precise and effective resource allocation. Additionally, extending the system's functionality to include temporal and contextual analysis could enable it to identify behavioral patterns over time, offering deeper insights into triggers like social isolation or bullying. These advancements would enhance the system's scalability and impact, ensuring proactive support for individuals at risk.

In conclusion, this study highlights the value of leveraging artificial intelligence to address critical mental health challenges. The model demonstrated a strong ability to detect suicidal ideation while maintaining a balance between precision and recall. While the findings are promising, future iterations must focus on reducing false negatives and incorporating advanced techniques to further refine the system's capabilities, ensuring it becomes an indispensable tool in suicide prevention efforts.

6.2. Future Directions

While the model demonstrated strong performance, further efforts are needed to address false negatives, as these represent critical missed opportunities for intervention. Incorporating additional data sources, expanding the dataset, and employing advanced techniques such as temporal analysis and topic modeling could improve the model's ability to identify context-specific risk factors. Additionally, embedding ethical considerations and privacy safeguards into the system ensures responsible deployment while maintaining user trust.

In conclusion, this study demonstrates the potential of AI-driven solutions to tackle the global mental health crisis through large-scale analysis of social media data. The model's ability to detect suicidal ideation effectively aligns with the objectives of scaling proactive suicide prevention and advancing data-driven models. Future enhancements will further refine its capabilities, paving the way for impactful applications in mental health intervention and resource allocation.

7. Conclusions

This study set out to address the global mental health crisis by leveraging artificial intelligence (AI) to detect suicidal ideation from social media posts. The primary goal was to develop a data-driven model capable of analyzing large-scale Twitter datasets to identify patterns indicative of distress. Through advanced natural language processing (NLP) and sentiment analysis techniques, the model aimed to provide timely and actionable insights to support suicide prevention efforts. By achieving a balance between precision and recall, the model demonstrates significant potential for real-world applications in identifying at-risk individuals and facilitating targeted interventions.

The process involved several critical steps to ensure the model's effectiveness. A curated subset of the original dataset of 20,000 tweets was used to minimize noise and focus on high-quality, relevant data. Preprocessing techniques, such as tokenization, stemming, and feature extraction through TF-IDF and Count Vectorization, transformed raw text into a machine-readable format. A Random Forest Classifier was selected for its robustness and ability to handle high-dimensional data, ensuring that the model could effectively capture linguistic and emotional patterns associated with suicidal ideation. These steps were integral in refining the model's ability to distinguish between "Potential Suicide Posts" and "Not Suicide Posts."

The findings revealed that the model achieved an overall accuracy of 85%, with a precision of 88% and a recall of 83% for "Potential Suicide Posts," reflecting its strong capability to identify high-risk cases while maintaining a low false-positive rate. The Precision-Recall curve, with an AUC score of 0.93, further validated its effectiveness in balancing precision and recall across varying thresholds. The confusion matrices for both training and test datasets highlighted areas for improvement, particularly in reducing false negatives, which remain critical for sensitive applications like suicide prevention.

Overall, the study demonstrates the potential of AI-driven solutions to address critical mental health challenges by analyzing social media data to detect suicidal ideation. While the model performed well in capturing high-risk cases and minimizing false alarms, further enhancements are needed to improve recall and address missed cases. Incorporating advanced architectures like transformers, temporal analysis, and broader datasets can further refine the system. This research underscores the importance of leveraging AI to create scalable and proactive tools for mental health intervention, offering hope for a more comprehensive approach to suicide prevention.

References

1. Abdulsalam, A., & Alhothali, A. (2024). Suicidal ideation detection on social media: A review of machine learning methods. *Social Network Analysis and Mining, 14*(1), 1-16.
2. Abraham, Z.K., & Sher, L. (2019). Adolescent suicide as a global public health issue. *International Journal of Adolescent Medicine and Health.* https://doi.org/10.1515/IJAMH-2017-0036
3. Abdurrahim, A., & Fudholi, D.H. (2024). Mental health prediction model on social media data using CNN-BiLSTM. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control.* https://doi.org/10.22219/kinetik.v9i1.1849
4. Abreu, T. de, & Martins, M. das G. T. (2022). A presença de ideação suicida em adolescentes e terapia cognitivo-comportamental na intervenção: Um estudo de campo. *Revista Ibero-Americana de Humanidades, Ciências e Educação.* https://doi.org/10.51891/rease.v8i5.5525
5. Adagale, S., & Gupta, P. (2024). Comprehensive analysis of text-based sentiment analysis using deep learning. *IEEE ICITEICS.* https://doi.org/10.1109/iciteics61368.2024.10624933
6. Balasubramanian, J., Koppa, K.B., Solanki, V., & Saxena, A.K. (2024). Suicide thoughts screening with social media using cat swarm-intelligent adaptive recurrent network. *Multidisciplinary Science Journal.* https://doi.org/10.31893/multiscience.2024ss0501
7. Berger, G., Casa, A.D., & Pauli, D. (2015). Suizidalität bei Adoleszenten – Prävention und Behandlung. *Therapeutische Umschau. Revue Thérapeutique.* https://doi.org/10.1024/0040-5930/A000728
8. Bhor, S., Bhor, S., Rani, B., Aniket, F., & Dube, Prof. D. (2024). Implementation of sentiment analysis using deep learning. *International Journal of Advanced Research in Science, Communication and Technology.* https://doi.org/10.48175/ijarsct-16935
9. Botta, A., Mohini, Pandey, Raju, A., & Balaji, Ch. (2024). Deep learning for sentiment analysis in financial markets. *IEEE ICRTCST.* https://doi.org/10.1109/icrtcst61793.2024.10578359
10. Cai, S., Jung, H., Liu, J., & Liu, W. (2024). Research on the applicability of suicide tweet detection algorithms. *Applied and Computational Engineering.* https://doi.org/10.54254/2755-2721/55/20241512
11. Chugh, S., Bansal, Y., Nagpal, R., Sakshi, ., Kaur, S., Saluja, S., Ahluwalia, B.K., & Sharma, S. (2024). The impact of social media on mental health. *Mental Health Insights.* https://doi.org/10.4018/979-8-3693-2215-4.ch022

12. Elsayed, N., Elsayed, Z., & Ozer, M. (2024). CautionSuicide: A deep learning-based approach for detecting suicidal ideation in chatbot conversation. *arXiv.Org*. https://doi.org/10.48550/arxiv.2401.01023
13. Favril, L., Yu, R., Geddes, J.R., & Fazel, S. (2023). Individual-level risk factors for suicide mortality in the general population: An umbrella review. *The Lancet. Public Health.* https://doi.org/10.1016/s2468-2667(23)00207-4
14. Grover, C., Huber, J., Brewer, M., Basu, A., & Large, M. (2023). Meta-analysis of clinical risk factors for suicide among people presenting to emergency departments. *Acta Psychiatrica Scandinavica.* https://doi.org/10.1111/acps.13620
15. Hase, Y.P., Karwar, P.B., Hingmire, S.N., & Gopale, B.V. (2024). Sentiment analysis using deep learning. *International Journal of Advanced Research in Science, Communication and Technology.* https://doi.org/10.48175/ijarsct-17478
16. Joinson, D., Davis, O., & Simpson, E. (2024). The dynamics of emotion expression on Twitter and mental health in a UK longitudinal study. *International Journal for Population Data Science.* https://doi.org/10.23889/ijpds.v9i4.2437
17. Kansal, M., Singh, P., Srivastava, P., Singhal, R., Deep, N., & Singh, A. (2024). Mental health monitoring in the digital age. *Mental Health Monitoring.* https://doi.org/10.4018/979-8-3693-2359-5.ch011
18. Liu, H.Y., & Qayyum, Z. (2023). Suicidal behaviors in children and adolescents: Synthesis of issues and solutions from global perspectives. *JAAC.* https://doi.org/10.1016/j.jaac.2023.07.374
19. Mahmud, S.A. (2023). Suicidal tweet dataset. *Kaggle.* Retrieved November 29, 2024, from https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset/code
20. Meier, A., & Reinecke, L. (2023). Social media and mental health. *Oxford University Press.* https://doi.org/10.1093/oso/9780197520536.003.0012
21. Mohamed, N. (2024). Investigating the impact of social media use on adolescent mental health. *OSF Preprints.* https://doi.org/10.31234/osf.io/fbr84
22. Mostardeiro, V.M.P., Somavilla, V.E., & Mocelin, G. (2022). Ideação suicida no contexto militar. *Conjeturas.* https://doi.org/10.53660/conj-s02-1139
23. Moulahi, B., Azé, J., & Bringay, S. (2017). Suivi et détection des idéations suicidaires dans les médias sociaux. *Social Media Insights.*
24. Muniyapillai, T., Kulothungan, K., Jai, N., CM, S.S.K., Godwyn, R., Shivashankari, S., Raje, S., Krishnakumar, S.P., Devi, S., & Suresh, S. (2024). Suicide and its risk factors – An ecological study. *Journal of Education and Health Promotion.* https://doi.org/10.4103/jehp.jehp_940_23
25. Environmental Systematic Analysis of Factors Associated with Adolescent Suicide Risk. (2024). *Korean Association for Learner-Centered Curriculum and Instruction.* https://doi.org/10.22251/jlcci.2024.24.13.123

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.